(Partially) Ranking Visualizations Using (Log) Weber's Law

Roy G. Biv, Ed Grimley, *Member, IEEE*, and Martha Stewart

Abstract—Models of human perception — sometimes called laws — are valuable tools for deriving concrete recommendations in visualization design. However, we must be careful to assess the explanatory power of such models before putting them into practice, particularly if they do not model variance in individual performance. We present a secondary analysis of data from Harrison *et al.* [1], investigating the relationship between the correlation of two variables (measured using Pearson's *r*) and the precision of people's estimates of that correlation using different visualizations (measured using *just-noticeable differences*). We begin with a linear model — similar to that of the original work — and walk through a series of refinements to this model. We first address problems of non-constant variance, presenting evidence that a log-linear — rather than linear — model better describes the relationship between just-noticeable differences and objective correlation. We then augment our model with censored regression to include all observations in the analysis, including outliers, data near ceilings and floors resulting from features of the data collection process, and entire visualizations originally dropped due to large numbers of data points worse than chance. Finally, we adopt a Bayesian variant of our model to derive a partial ranking of visualizations of correlation based on the expected precision of people's estimates of correlation on a randomly-drawn dataset. This partial ranking provides concrete guidance to practioners by grouping visualizations with similar performance and by giving precise estimates of the difference in performance between groups of visualizations. We discuss the applicability of similar models to other problems of estimating the perceptual performance of visualizations from experimental data.

Index Terms—TBD

Introduction

TBD, motivation, blah blah blah. ….

The original paper used a modelling approach that removed individual variation before fitting the model: it fit a linear regression to the means of the just-noticeable differences within each visandsign * r * approach, not to the individual observations directly. By removing a large portion of the variance in the data (individual differences), they could not use their parametric model to make predictive inferences. Instead, they employed non-parametric tests to examine differences between visualization types, which complicates the estimation of effect sizes. Even if we establish that one visualization is better than another, we would like to know by how much in order to judge whether the difference is meaningful in practice. Ideally we would like to know this effect size on some interpretable scale (such as in terms of just-noticeable differences in r), which is made difficult when using non-parametric tests.

Much of this paper focusses on understanding and accounting for individuals' differences in precision of estimation in order to derive parametric models that can predict the expected precision of each visualization technique. Given an appropriate parametric model, we can estimate interpretable differences between visualization types (e.g., as a ratio of just-noticeable differences) in order to judge whether these differences have practical significance. To derive such a model, we first tweak the original linear model to fit it to individual observations. Through a series of refinements, we construct a censored, log-linear regression model that improves on the fit of the linear model and accounts for individual differences *and* artifacts of the experimental design.

This model allows us to directly and quantitatively answer questions left largely unaddressed by the original paper: given a dataset with unknown correlation, how well would we expect each visualization technique to perform? What are the practical differences in performance? Which visualizations are effectively equivalent? We identify clusters of visualizations with similar precision and quantify the expected difference in precision between clusters, yielding a comprehensive set of practical recommendations in the form of a partial ranking of visualizations of correlation.

In the rest of this paper, we briefly overview the original experimental design and model from Harrison *et al.* [1]. We then walk through a series of models: a linear model based on the original paper, a log-linear model addressing problems of non-constant variance and skewed residuals, a censored model addressing floors and ceilings in the data caused by artifacts of the experiment, and a Bayesian model that can estimate the performance of each visualization on a class of datasets. From this last model we derive a partial ranking of visualizations of correlation.

1 Background

- Should briefly describe original paper's experimental setup
- Might be some useful pieces about perceptual laws to put in here. Might also be out of scope for this.

2 Model 1: Linear Model

- Why individual observations? No individual observations => no good estimate of precision, so we have no idea how good estimates of effect size were in original paper even if we could derive them
- Tweaking the original model so we can include individual observations: addressing the approach bias; similarity of results using tweaked model to original paper's model.
- Problems with linear model:
  - No floor at 0
  - Non-constant variance
  - Skewed residuals

- *Roy G. Biv is with Starbucks Research. E-mail: roy.g.biv@aol.com.*
- *Ed Grimley is with Grimley Widgets, Inc.. E-mail: ed.grimley@aol.com.*
- *Martha Stewart is with Martha Stewart Enterprises at Microsoft Research. E-mail: Martha.stewart@marthastewart.com.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send *e-mail to: tvcg@computer.org.*

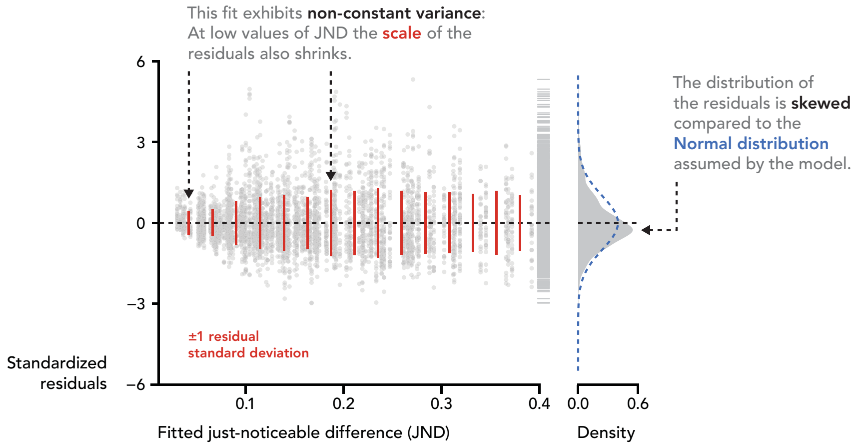

Fig. 1 This will be a comparison of linear and log-linear model fits (currently just linear fit shown)

3 Model 2: Log-Linear Model

- Addresses floor, non-constant variance, skewed residuals
- Re-examine cutoffs from original paper:
    - MAD identification of "outliers"; side note that this did not solve non-normality (but we have now solved it).
    - P(chance) threshold: chucking whole visualizations only solves half the problem, you still have to deal with individual observations greater than chance in visualizations you *do* analyze. Censoring is good candidate here
    - Ceiling in JND (when approach is from above) and floor in JND (when from below) also candidates for censoring. This point may belong in the next section.

4 Model 3: Censored Log-Linear Model

- Derive censoring thresholds
- Include example of censored distribution as explanation?
- Show example of downward bias at low r in a viz near the chance threshold to demonstrate value of censored models

The problem is that we have excluded certain visualizations for having too many observations worse than chance, but have done nothing to address those observations worse than chance in the visualizations we did analyze. Importantly, in the case of points near or beyond this boundary, we can say that these observations probably represent JNDs of .45 or worse, but that we do not actually know the exact JND due to the constraints of the experiment. This type of data can be analyzed using censored regression.

Censored regression is used when some of the observed data points do not have a known value, but instead are known to lie above (or below) a certain threshold. While we do not know the exact value of points beyond the threshold, we still know *how many* points were observed beyond the threshold, and it is this information that we can use to fit the model. In this case, while we cannot observe certain values of JND — either because the setup of the experiment makes them indistinguishable from chance, or because of ceilings and floors in observable JND due to the bounds on r — we can use observations close to or beyond those thresholds to estimate the proportion of values we might expect to see above them.

5 Model 4: Bayesian Censored Log-Linear Model

- Derive priors
- Demonstrate model
- Possibly add participant effect
- Simulate drawing random distribution to derive expected differences given unknown correlation

6 Partial Ranking of Visualizations

- Given a problem space with datasets having some known/estimated distribution of r, easy to re-compute rankings from the model (possibly put in discussion)

7 Discussion

- Censoring probably useful in other similar experiments with thresholds that are artifacts of the data collection process.
- Might be something in here about perceptual "laws", model fitting, individual differences, etc

8 Conclusion

Acknowledgments

- Original authors?

References

[1] L. Harrison, F. Yang, S. Franconeri, and R. Chang, "Ranking Visualizations of Correlation Using Weber's Law," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1943–1952, 2014.